bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

An Integrated Workflow for Enhanced Taxonomic and Functional Coverage of the Mouse Faecal Metaproteome

Nicolas Nalpas[1], Lesley Hoyles[2,3], Viktoria Anselm[1], Tariq Ganief[1], Laura Martinez-Gili[2], Cristina Grau[4], Irina Droste-Borel[1], Laetitia Davidovic[5], Xavier Altafaj[4,6], Marc-Emmanuel Dumas[2,7,8], Boris Macek[1]

[1] Proteome Center Tuebingen, University of Tuebingen, Germany; [2] Biomolecular Medicine Section, Division of systems Medicine, Department of Metabolism, Digestion and Reproduction, Imperial College London, Sir Alexander Fleming building, London SW7 2AZ, UK; [3] Department of Biosciences, Nottingham Trent University, UK; [4] Bellvitge Biomedical Research Institute, Spain; [5] Université Côte d'Azur, CNRS, Inserm, IPMC, France; [6] Neurophysiology Unit, University of Barcelona - IDIBAPS, Spain; [7] Genomic and Environmental Medicine, National Heart & Lung Institute, Faculty of Medicine, Imperial College London, London, SW3 6KY, United Kingdom. [8] European Genomic Institute for Diabetes, INSERM UMR 1283, CNRS UMR 8199, Institut Pasteur de Lille, Lille University Hospital, University of Lille, 59045 Lille, France.

Short title: Enhanced workflow for metaproteomics.

Keywords: Metaproteomics; Microbiome; Mus musculus; Mass spectrometry; Proteogenomics.

*To whom correspondence should be addressed:
Prof. Dr. Boris Macek
Proteome Center Tuebingen
Interfaculty Institute for Cell Biology
Auf der Morgenstelle 15
72076 Tuebingen
Germany
Phone: +49/(0)7071/29-70558
Fax: +49/(0)7071/29-5779
E-Mail: boris.macek@uni-tuebingen.de

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Abstract

The intestinal microbiota plays a key role in shaping host homeostasis by regulating metabolism, immune responses and behaviour. Its dysregulation has been associated with metabolic, immune and neuropsychiatric disorders and is accompanied by changes in bacterial metabolic regulation. Although proteomics is well suited for analysis of individual microbes, metaproteomics of faecal samples is challenging due to the physical structure of the sample, presence of contaminating host proteins and coexistence of hundreds of species. Furthermore, there is a lack of consensus regarding preparation of faecal samples, as well as downstream bioinformatic analyses following metaproteomic data acquisition. Here we assess sample preparation and data analysis strategies applied to mouse faeces in a typical LC-MS/MS metaproteomic experiment. We show that low speed centrifugation (LSC) of faecal samples leads to high protein identification rates and a balanced taxonomic representation. During database search, protein sequence databases derived from matched mouse faecal metagenomes provided up to four times more MS/MS identifications compared to other database construction strategies, while a two-step database search strategy led to accumulation of false positive protein identifications. Comparison of matching metaproteome and metagenome data revealed a positive correlation between protein and gene abundances, as well as significant overlap and correlation in taxonomic representation. Notably, nearly all functional categories of detected protein groups were differentially abundant in the metaproteome compared to what would be expected from the metagenome, highlighting the need to perform metaproteomics when studying complex microbiome samples.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Introduction

The prokaryotic component of the gut microbiota has multiple roles, contributing to carbohydrate fermentation, maintenance of gut barrier integrity, as well as antimicrobial and immunomodulation activities [1,2]. In metabolically healthy humans and mice, the gut microbiota is predominated by two to three bacterial enterotypes [3-5]. These enterotypes display significant heterogeneity in terms of species number, composition and relative abundances depending on the location of the sample (upper vs lower GI tract) or the timing (circadian variations) [6,7]. The gut microbiome has recently been associated in a number of conditions ranging from inflammatory bowel syndrome to Parkinson's disease [8-11]. Interestingly, an increasing number of studies have reported a correlation between the gut microbiome and neurodevelopmental disorders [12-14]. Notably, this includes changes in the gut microbiome of Down syndrome individuals in comparison to non-trisomic individuals [15]. Given the established interaction between the host and the gut microbiome, a functional analysis of the gut microbiome may help in understanding its contribution to pathophysiology. To study the gut microbiome, approaches relying on nucleotide sequencing have so far been preferred by the scientific community due to lower experimental costs, higher data throughput and proven analytical workflows. However, metagenomics can only assess the genetic potential, whereas metaproteomic investigates produced proteins (and therefore functions). In particular, microbiome functional analysis can be performed using high-resolution mass spectrometry (MS), to measure either protein abundance or metabolite production [16-18]. Although bacterial MS-based proteomic is well established, metaproteomic sample preparation is hindered by many challenges, such as physical structure of the sample, the presence of contaminating proteins and the coexistence of hundreds of microorganisms. Several studies in human have shown that different sample preparation methodologies can result in significant changes in the taxonomic composition and functional activities representation [19,20]. Beyond

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

sample preparation, the bioinformatic processing of metaproteomic data remains challenging, due to the choice of representative protein sequence database, the redundancy in protein functional annotation and elevated false discovery rate for peptide identification. Some of these challenges have already been addressed by published software packages, such as MetaProteomeAnalyzer [21] and MetaLab [22], which are all-in-one metaproteomic analytical workflows, or UniPept [23], which allows peptide-based taxonomic representation.

However, a number of bottlenecks remain unaddressed, namely: (1) the lack of appropriate sample preparation methods for optimal protein identification and taxonomic representation, (2) high false positive rates in searches involving very large databases, (3) the impact of protein sequence database constructions on protein identification, (4) the shortcomings of taxonomic representation derived from MS-based peptide identification and (5) the lack of unbiased assessment of the functional enrichment provided by the metaproteome compared to its matching metagenome potential.

Here, we present a state-of-the-art LC-MS/MS-based workflow for the optimal metaproteome characterisation of murine faecal samples. In terms of sample preparation, we achieve the highest protein identification and biological correlation when combining LSC with in-solution digestion. In terms of data processing, we demonstrate that the false discovery rate can only be controlled using a single-step database search and that protein sequence database derived from matching metagenome provides superior identification rate compared to publicly available databases. We show that accurate taxonomic representation can be derived from peptide spectral match (PSM) count. And while we overall observed a positive correlation between protein and gene abundances, the metaproteome showed an enrichment in functionally-active pathways compared to matching metagenomic potential.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Results

Low-speed centrifugation increases peptide identification rates

Our initial experiment involved the establishment of an optimal sample preparation workflow applied to the mouse faecal metaproteome. In this context, we assessed several critical steps within our sample preparation method: 1) the usage of LSC; 2) in-solution digestion versus filter-aided sample preparation (FASP); and 3) cell lysis by bead-beating, nitrogen or sonication (**Figure S1A**).

The number of peptides identified per MS raw file in the LSC group was significantly higher with nearly 23 % more peptide identifications (**Figure 1A**). This was also observed at the protein group level, but to a lower extent. Approximately 15 % of protein groups were identified by a single peptide, while the median protein sequence coverage was 18.7 %. Such metrics are usually indicative of highly complex samples that are not completely covered by a single MS measurement under the stated parameters.

In-solution digestion consistently outperformed FASP based on peptide and protein groups identification (**Figure 1B**). Compared to other methods, in-solution digestion combined with LSC procedure provided nearly twice as many peptide identifications and 30 % more protein groups. Furthermore, there was much less variability in the number of peptides and protein groups identified across samples with this method.

When assessing the different lysis methods (in combination with LSC procedure), the methods employed showed similar efficacies to identify peptides or protein groups (**Figure 1C**). This suggests that any of these lysis methods could be used for preparation of mouse faecal samples. However, the time necessary for the bead-beating approach was 30 min, while the nitrogen grinding approach required 150 min and the sonication approach needed 90 min. The bead-

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

beating method would therefore provide a better time optimisation, particularly for large-scale metaproteomic experiments.

LSC aids in recovery of Bacteroidetes proteins, whereas nLSC favours Firmicutes proteins

Peptides that were identified after LSC and nLSC, were analysed to identify their phylogenetic origin. The lowest common ancestor was determined using the Unipept interface [24], which assigns peptide sequences to operational taxonomic units (OTUs). Based on the PSM count, the most abundant superkingdom was bacteria, representing 77.7 % and 82.2 % of all PSMs for LSC and nLSC, respectively (**Figure 1D-E**, **Table S1**). The second most represented superkingdom was eukaryota in both the LSC and nLSC procedures. Under the assumption that eukaryotic peptide sequences originated from the host, the proportion of *Mus musculus* proteins was investigated further using intensity-based absolute quantification (iBAQ) values. The LSC samples contained on average nearly two-fold more murine proteins (20.4 %) in comparison to nLSC samples (14.6 %) (**Figure S1B**). Such findings were surprising since the use of the LSC method was reported in a previous study to help with the removal of human cells [19]. We also investigated the presence of peptides from host diet and found very low levels of dietary peptides contamination (approximately 2 %), which was higher among LSC-prepared samples (**Figure S1C**). This suggests that the majority of dietary proteins are absent or depleted during the initial solubilisation step of the faecal pellet, a step common to both procedures.

At the phylum level, three main taxa were represented in both LSC and nLSC: *Bacteroidetes*, *Firmicutes* and *Metazoa*. There were large differences in the number of PSMs assigned to the two main bacterial phyla when comparing LSC and nLSC methods. *Bacteroidetes* accounted for 33 % and 13 % of PSMs, whereas Firmicutes amounted to 24 % and 48 % of PSMs in LSC and nLSC procedures, respectively. The PSM counts were also calculated for each individual

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

sample across the two procedures in order to investigate the proportion of *Firmicutes* and *Bacteroidetes*. We observed a significant inversion in the *Firmicutes*:*Bacteroidetes* ratio between LSC and nLSC (**Figure S1D**). Interestingly, large sample-to-sample variation in the *Firmicutes* to *Bacteroidetes* ratio was observed in the nLSC group. Such variation seemed biological rather than technical, since technical replicates (samples measured multiple times by MS) displayed relatively similar ratios. Overall, our results suggest different bacterial representation, whereby LSC favours *Bacteroidetes* and corresponds more to genome-derived taxonomy.

LSC and nLSC methods are characterised by different protein abundance profiles

We further investigated the overlap between the peptides or protein groups identified following either LSC and nLSC procedures (**Figure 2A**). In terms of peptides, only 27.7 % were identified with both procedures, the rest of the peptides being split equally into unique to LSC and nLSC methods. Similar results were observed at the protein groups level with 38.7 % of protein groups being identified in both procedures. Label-free quantitative (LFQ) comparison between LSC and nLSC procedures revealed an intermediate correlation ($\rho = 0.44$) (**Figure S2A**). Notably, the correlation between technical replicates was very high. This was illustrated further through a principal component analysis (PCA) (**Figure 2B**). Clustering of technical replicates confirmed the high technical reproducibility. Our findings indicate that while the two procedures have a poor identification overlap and quantification correlation, the main differences may still result from biological variations.

Using LFQ intensities, we then performed a *t*-test to identify which protein groups have different abundances between the two procedures. Out of 2,589 quantified protein groups, 365 and 267 were significantly up-regulated and down-regulated in LSC versus nLSC samples, respectively (FDR ≤ 0.01 and absolute fold-change ≥ 2.5) (**Figure 2C**, **Table S2**). We gained functional insights into these differences by performing an over-representation analysis of

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

KEGG pathways. The over-represented pathways based on the up- or down-regulated protein groups were mostly similar (FDR ≤ 0.1) and were associated with core microbial functions, such as ribosome, carbon metabolism and carbon fixation pathways (**Figure 2D**, **Table S2**). The protein groups unique to LSC or nLSC showed over-representation of protein export in the LSC samples, whereas biosynthesis of amino acid, fatty acid degradation and bacterial chemotaxis were over-represented in the nLSC samples (**Figure S2B**). Mapping of the individual protein groups on the glycolysis-gluconeogenesis KEGG pathway highlighted the redundancy in protein function (i.e. protein with identical function found across multiple OTUs) and discrepancy in abundance (i.e. OTU depletion leads to reduced protein abundance, while OTU enrichment increases protein abundance) (**Figure S2C**). Protein differential abundance testing confirmed the divergence between LSC and nLSC procedures and was suggestive of broad taxonomic changes, rather than variation in functional activities.

MS instrument selection critically determines the identification rate

Following assessment of sample preparation, we investigated the identification rates obtained through LC-MS/MS measurements with two different mass spectrometers, namely Orbitrap Elite and Q Exactive HF. In this context, we prepared samples using the LSC method from faeces collected in a cohort of 38 mice. The newer generation Orbitrap instrument, namely Q Exactive HF, provided a median of 229 MS/MS spectra identification per minute, while the Orbitrap Elite resulted in less than half that number (**Figure S3A-B**). A similar significant trend was also observed at the peptide and protein group levels, despite the fact that the measurement time was halved on the Q Exactive HF versus the Orbitrap Elite instrument. Of note, measurement on the Q Exactive HF required half of the sample material as compared to the Orbitrap Elite (0.5 µg/h vs 1 µg/h), a critical point when processing limited amount of material. While expected, our findings highlight the crucial impact of mass spectrometer speed and

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

sensitivity in a typical metaproteomic measurements. Indeed, the choice of MS instrument was among the parameters with the greatest impact on identification rates.

Two-step database search strategy shows a dramatic increase in false positive rate

After acquisition of LC-MS/MS raw data, the MS/MS spectra are searched against a protein sequence database. One aspect of database search is the controversial use of a two-step search strategy, whereby LC-MS/MS measurements are processed initially against a large protein sequence database with no FDR control (FDR $\leq$ 1). Subsequently, the original database is filtered to retain only protein sequences that were identified during the first search. During the second database search, the measurements are processed against the reduced database with FDR control (e.g., FDR $\leq$ 0.01) [25]. To assess the false discovery rates in such approach, we searched a single HeLa cell LC-MS/MS file using MaxQuant software against a *Homo sapiens* protein sequence database supplemented with different number of bacterial protein sequences (**Figure S3C**). The HeLa measurement is used here as a proxy for a complex microbiome measurement, with the exception that the sample composition is known (i.e. of human origin). We initially established a gold standard by processing the HeLa measurement only against an *H. sapiens* database, which resulted in approximately 5,000 human (eukaryota) protein groups identified for the single-step search at FDR $\leq$ 0.01 (**Figure 3A**, **Table S3**). Notably, the same database used in a two-step search identified less than 1 % additional protein groups in comparison to a single-step search, despite nearly twice as much processing time. We then processed our HeLa measurement against *H. sapiens* database supplemented with 1:1, 1:2, 1:5, 1:10 and 1:20 *H. sapiens*:bacteria protein sequences, resulting in increasingly large databases (**Figure S3C**, **Table S3**). For the single-step database search against the 1:20 database, we observed a 10 % decline in the number of human protein groups identified, while 132 bacterial protein groups were identified (false positives). On the contrary, the 1:20 two-step database search resulted only in a 1 % decrease compared to the gold standard. This processing also

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

231 revealed a large number of bacterial protein groups identification (980 protein groups).
232 Furthermore, the two-step search led to large number of MS/MS spectra to be assigned to
233 different sequences (or newly assigned) in comparison to the gold standard (**Figure S3D**, **Table**
234 **S3**); this phenomenon was much less pronounced when performing the single-step search.
235 We then calculated the actual FDR for each processing approach using either the reverse hits
236 or the reverse hits plus the bacterial hits (which in our case are false positives). For both the
237 single-step and the two-step search, we obtained an FDR of 2.6 % when using only the reverse
238 hits for FDR calculation (**Figure 3B**). However, when using the reverse hits plus the bacterial
239 hits, we calculated an actual FDR of 8 % and 34 % for the single- and two-step search with
240 1:20 database, respectively. This represents a dramatic increase in the rate of false positive
241 identification when using two-step search, despite controlling for 1 % FDR. Notably, these
242 false positive hits would remain unnoticed in a microbiome sample of unknown composition,
243 thus highlighting the inherent problem associated with the two-step database search.

244 Metagenome-derived protein sequence database outperforms databases from public
245 resources

246 We then assessed the importance of the protein sequence databases and their impact on the
247 identification of microbiome sample proteins. In this context, we used faeces collected in a
248 cohort of 38 mice that were measured on Q Exactive HF instrument. We compared four
249 different databases, namely UniProt bacterial reference proteomes (5,408,622 protein entries),
250 UniProt bacterial pan proteomes (18,541,701 entries), protein sequences from the Mouse Gut
251 Metagenome catalogue (2,626,630 entries) [26] and protein sequences obtained from our
252 matched-metagenome data (1,595,268 entries) (**Figure 3C**). Notably, the number of protein
253 groups identified was inversely proportional to the number of protein sequences in the
254 databases, with the matched-metagenome database resulting in 25,230 identified protein
255 groups—that is seven-times more than the UniProt reference proteome database. A similar

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

trend could be observed at the peptide and MS/MS levels, with on average 23.23 % MS/MS spectra identified using the matched-metagenome database. Not surprisingly, the matched-metagenome database yielded the most identifications; however, the murine microbiome catalogue also performed surprisingly well, making it an excellent substitute in cases where the matched metagenome is not available.

Taxonomic representation correlates significantly between metaproteome and metagenome down to species level

We also investigated the correlation in taxonomic representation between the metagenomic and metaproteomic datasets. Because metaproteomics is not generally used as the method of choice to infer taxonomy in a sample, there are fewer software tools available for this purpose in comparison to metagenomic approaches. Initially, we retrieved taxonomic assignment using Kraken2 software [27] and compared six methods to calculate taxon abundance using our metaproteome data for our cohort of 38 mice (**Figure 4A**, **Table S4**). A range of filtering thresholds were also implemented to remove low abundant taxa. The Kraken2 software in combination with PSM count resulted in high OTU abundance correlation between metagenome and metaproteome datasets even without any taxonomic representation filtering. This correlation increased when a minimum threshold of 0.4 % total taxonomic representation was used; however, it also reduced the number of identified OTUs by more than half. Aside from the comparison to the metagenome, we investigated the taxonomic assignment obtained using Diamond, Unipept and Kraken2 software tools [23,27,28]. This revealed far superior results when using Kraken2, with between 20 and 30 % PSM assigned to a species, as opposed to approximately 5 % with Diamond or Unipept (**Figure S4A**).

Using Kraken, we inspected the lowest taxonomic level that could be reached while still displaying significant correlation against metagenomes (**Figure 4B**). Highly significant correlations in taxonomic abundances were observed at all taxonomic levels. At the phylum

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

level, the ratio of *Firmicutes* to *Bacteroidetes* (or *Bacteroidota*) was significantly different between metagenomes and metaproteomes, nevertheless ratios were consistent between the two omics approaches and ranged between 0 and 1 (**Figure S4B**). When assessing the identification of different phyla between omics datasets, we did not observe technical bias relating to sample preparation or MS sensitivity (**Figure S4C**). Indeed, the low number of identifications in metagenomic data did not necessarily translate into reduced or missing identification in metaproteomic data; in addition, Gram-positive phyla were not necessarily over-represented among the missing phyla in the metaproteomes. Under the current conditions, metaproteomic can be used to assess taxonomic biomass [29] from phylum down to species level for the mouse faecal microbiota.

Metaproteome to metagenome correlation highlights an over-representation in the core microbiome functions

Due to the availability of matching metagenomic and metaproteomic data for our cohort of 38 mice, we assessed the correlation between gene and protein abundances. To deal with the intrinsic difference between the two datasets, the gene entries were grouped in a similar fashion as the protein groups (i.e. based on peptide identification) and the maximum expression was calculated per gene group. Here, we show that a majority of gene-protein pairs (91 %) have a positive correlation, with a median of 0.39, the rest having a median negative correlation of -0.09 (**Figure 4C**, **Table S4**). Notably, 3,519 gene-protein pairs displayed a significant positive correlation.

To identify the core pathways within our mice cohort, we performed an over-representation analysis of the significantly correlated gene-protein pairs (**Figure 4D**, **Table S4**). Among these pairs, there was an over-representation in carbon fixation, glycolysis-gluconeogenesis, citrate cycle and carbon metabolism pathways (KEGG) [30]. We further characterised the correlating genes and proteins based on gene ontology (GO) and identified 178 biological processes

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(GOBP), 51 cellular components (GOCC) and 20 molecular functions (GOMF) that were over-represented (**Figure S4D-F**, **Table S4**). Our results confirm the central role of carbon fixation and general metabolism, which are associated with bacterial energy production, in the murine faecal microbiome under the analysed conditions.

The metaproteome is enriched in functionally active pathways compared to the matching potential encoded in the metagenome

The metagenome corresponds to the microbiome genetic potential, whereas the metaproteome represents its truly expressed functional activities. Thereby, we compared the functional abundance derived from the metagenomic versus metaproteomic datasets within our cohort of 38 mice. To allow comparison, the KEGG level 2 categories were quantified and normalised separately for each omics datasets (**Figure S5A**, **Table S5**). Out of 55 KEGG categories, we found 15 and 37 to be significantly increased and decreased in abundance at the metaproteome level in comparison to the metagenome (FDR $\leq$ 0.05). In general, the metagenome-based quantification of KEGG categories was stable across categories, whereas large differences were observed for the metaproteome.

To prioritise the KEGG categories, we selected eight categories differing significantly in terms of gene-protein correlation in comparison to the overall correlation (**Figure 5A** and **S5B**). Among the KEGG categories displaying higher abundance in the metaproteome compared to the metagenome were the membrane transport, translation, signalling and cellular processes, and genetic information processing. Conversely, transcription, carbohydrate metabolism and antimicrobial drug resistance exhibited lower abundance. The KEGG Orthology (KO) entries differing significantly in abundance between the metagenomes and metaproteomes were identified via *t*-test and used for gene set enrichment analysis (GSEA). GSEA revealed an enrichment of a number of overlapping KEGG pathways, with 19 and 6 pathways positively and negatively enriched, respectively (**Figure 5B**, **Table S5**). Interestingly, we found the

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

ribosome pathway enriched in protein with increased abundance (between metaproteome and metagenome datasets), therefore highlighting the functional activation of this pathway (**Figure 5C** and **S5C**). Conversely, homologous recombination, DNA replication and mismatch repair were enriched in protein with decreased abundance, suggesting no or low activation of these pathways. Overall, our findings highlight the critical importance of metaproteomic to characterise microbiome samples particularly when it comes to their functional activity.

Discussion

Here, we provide solutions to some key bottlenecks hindering metaproteomic of murine faecal samples in order to enhance protein identification, taxonomic and functional coverage. These solutions include (1) an adequate sample preparation method, (2) the best strategy to control for false positive rates, (3) the ideal protein sequence database construction, (4) an accurate MS-derived taxonomic representation and (5) the leverage provided by metaproteomic to determine functionally enriched pathways.

An integrated workflow that provide the highest identification rate for metaproteomic of murine faecal samples and is amenable to other hosts

To the best of our knowledge this is one of the largest and most extensive comparisons undertaken to date, comprising over 40 different biological samples and over 200 LC-MS/MS runs. Overall, we reached identification rates that are similar to bacterial shotgun proteomics (ca. 20-40 %). In comparison to previous murine faecal metaproteomic studies, we identified more non-redundant peptides per samples (approximately 20,000 non-redundant peptides on a 60 min gradient) [31,32]. Several parameters may have influenced our greater performance, among which are the use of a faster and more sensitive Orbitrap instrument (i.e. Q Exactive

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

HF) [33,34], an optimised LC gradient [35] and a more representative protein sequence database (i.e. mouse metagenome catalogue or mouse matching metagenome) [26].

It should be noted that a number of aspects detailed herein would be directly applicable to metaproteomic study of human samples. Indeed, many of our conclusions are not connected to the taxonomic composition of species-specific samples and should therefore be transferrable [36]. Regarding the sample preparation, while the choice of LSC was in part based on taxonomic representation that is host-specific, the superior identification rate still encourages its usage in other host organisms. Noteworthy, the selected approaches used to control false discovery rate, construct protein sequence database and derive peptide-based taxonomic representation will likely hold true in many diverse metaproteomic samples.

Increased identifications are obtained when using LSC with in-solution digestion

Our study confirms previous observation with regard to the depletion or enrichment of several major bacterial phyla, which is dependent on laboratory preparation method and specifically the usage of differential centrifugation [19]. In this context our results do not match with the study from Tanca and colleagues, who reached opposite conclusions. However, there are several possible explanations for such discrepancy, such as the host organism under study (*i.e.* *Mus musculus* versus *Homo sapiens*) and different protein sequence database construction (*i.e.* mouse microbiome catalogue versus UniProtKB custom microbiome). Nonetheless, to preserve the trend in phylum distribution observed at the metagenome level in *Mus musculus*, the LSC approach seems more appropriate based on previous studies, as well as our direct comparison of the *Firmicutes* to *Bacteroidetes* proportion between metaproteomes and matching metagenomes [37,38]. The LSC approach also leads to more consistent identifications and as a result fewer missing values, which is a general and extensive problem in metaproteomic datasets. Therefore, we recommend the use of LSC when preparing murine faecal samples for measurement by mass spectrometry (Table 1).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

While we observed a significant increase in identification when using in-solution digestion in comparison to FASP, the lysis methods using either bead-beating, nitrogen grinding or sonication did not impact performance. Our results partially confirm the study from Zhang and colleagues regarding the superior performance of in-solution digestion [20] as opposed to other protein digestion strategies [39], possibly due to limited protein loss.

Here, we identified numerous protein groups showing significant changes in abundance between the LSC and nLSC approaches, as well as the over-representation of key KEGG pathways. Notably, degradation of carbohydrates and proteins is over-represented in LSC, while biosynthesis of amino acids is characteristic of the nLSC approach. While, these observations show an opposite trend compared to the study by Tanca and colleagues [19], possibly due to difference in host organism (*i.e. M. musculus* versus *H. sapiens*). Our results are similarly indicative of broad taxonomic changes more so than variation in functional activities.

Single-step search against matching-metagenome protein sequence database allows control of false discovery rate and highest protein identification

Currently, many metaproteomic studies use two-step database searches as ways to boost identification rates [25]. However, we demonstrate that this type of search dramatically underrepresents the number of false positives, due to the use of a decoy search strategy that is unsuitable in this context. Our results elaborate on a previous study by Muth and co-workers, who also emphasised the drawbacks of using a two-step search together with decoy strategy [40]. Here, our findings were so extreme that the number of false positives was equal or greater to the number of false negatives, with FDR outside of the accepted range (i.e. FDR > 0.1). We argue that the use of a two-step search should be avoided whenever possible and replaced by alternative strategies, such as taxonomic foreknowledge or using matching metagenomes (Table 1) [41].

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

We also show that the choice of protein sequence database had a serious impact on the identification rates, leading to nearly five-fold differences. While the best results were obtained using protein sequences derived from matched metagenomes, it was surprising that concatenation of bacterial protein sequences from online resources performed so poorly. This highlights the large diversity in proteins and peptides, which are not accounted among well characterised bacterial species, such as sequences from UniProtKB Reference or Pan proteomes. However, our results also reveal the importance of microbiome characterisation studies performed in different organisms (e.g. *H. sapiens*, *M. musculus*) or in specific tissues (e.g. oral, nasal) [26,42,43]. Indeed, the *M. musculus* microbiome catalogue led to identification rates similar to our matched metagenomes without the associated costs. In the future, more comprehensive microbiome catalogues may completely alleviate the need for matched metagenome (Table 1).

Accurate taxonomic annotation and quantification are obtained via Kraken2 software and PSM count

Based on the approaches tested here and in other studies, it is possible to derive taxonomic representation and abundance from MS-based peptide identification [29,44]. In this study, taxonomic representation correlated significantly between metaproteomic and metagenomic data at all taxonomic levels tested (*i.e.* phylum down to species), thus confirming observations from Erickson and colleagues [45]. Interestingly, the metaproteome PSM count is the calculation method that is closest to metagenome read count, it is therefore unsurprising that this calculation method showed the best correlation for taxonomic representation between the two omics datasets. While the Kraken2 software [27] provided the best taxonomic annotation (Table 1), it should be mentioned that the comparison to Diamond or Unipept softwares [23,28] is not entirely fair, since the metagenome taxonomy was solely derived from Kraken2 and may thus result in a favouring bias. Importantly, the proportion of *Firmicutes* and *Bacteroidetes*

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

displayed a similar trend (despite a significant difference) between omics, indicating that overall murine faecal microbiota contains more *Bacteroidetes* than *Firmicutes* [37,38]. While, the majority of OTUs were commonly identified across omics datasets (especially at higher taxonomic levels), there were several OTUs quantified exclusively at the metaproteome or metagenome level, e.g. *Firmicutes_H, Euryarchaeota, Thermoplasmatota*. Some of these discrepancies might be explained by different analytical artefacts (i.e. different sensitivity between Thermo Orbitrap versus Illumina HiSeq instruments). Yet it is important to state that bacterial activity and bacterial presence are different, therefore it is unsurprising to report only a medium overlap between metaproteome and metagenome [29].

The metaproteome shows an enrichment in functionally-active pathways compared to the matching metagenomic potential

Here, we observed an overall positive correlation between gene and protein abundances derived from metaproteome and matching-metagenome analysis. This was previously reported in a longitudinal study of metaproteome/metagenome fluctuations from one individual with Crohn's Disease [46]. In our case the significantly correlated entries were associated with core bacterial metabolic functions, such as carbon and energy metabolism or electron transfer activity [47]. Despite such correlations, we also reported extensive differences in quantified functions between metagenomic and metaproteomic. Notably, with regard to genetic information processing (KEGG level 2), the ribosome pathway was over-represented in entries with higher abundance in metaproteomes, whereas pathways associated with DNA repair, replication or recombination were over-represented in entries with increased abundance in metagenomes. This greatly highlights the main advantage of metaproteomic, which capture functionally active pathways, as opposed to the genetic potential represented by metagenomic [48]. Thus, these approaches are complementary to each other and can provide a more comprehensive understanding of a biological system.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Conclusion

To conclude, in this study we present an integrated analytical and bioinformatic workflow to improve protein identification, taxonomic and functional coverage of the murine faecal metaproteome. Notably, this workflow should be easily amenable to studying the human faecal metaproteome. LSC combined with in-solution digestion provided the highest identification rates. We also show that fast and accurate MS data processing can be achieved using a single-step database search against publicly available metagenome catalogues or matching metagenomes. Taxonomic representation can be generated directly from MS-based peptide identification. While protein and gene abundances show an overall positive correlation, the metaproteome showed a significant functional enrichment compared to its metagenomic potential; thus, emphasizing the need for more metaproteomic studies for adequate functional characterisation of the microbiome.

Methods

Animals and faecal samples collection

Mouse faecal pellets obtained from a small cohort of six wild-type B6EiC3SnF1/J mice were used to compare sample purification and protein extraction methodologies (**Figure S1A**). A larger cohort of 38 mice (euploid and trisomic Ts65Dn) was used to obtain mouse faeces, for further sample preparation described below (**Figure S1A**). Mice were housed and faeces were collected following the experimental procedures evaluated by the local Ethical Committee (Barcelona Biomedical Research Park, Spain). After collection, faecal pellets were frozen and stored at -80 °C until analysis.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

DNA extraction and whole-genome sequencing

Whole genome analysis was performed on the mouse cohort used for data analysis assessment. In brief, DNA was extracted from faecal samples using the FastDNA SPIN Kit (MP Biochemicals) and following manufacturer's instructions. DNA concentration was measured using a Qubit fluorometer (Invitrogen) and samples were shipped frozen to the Quantitative Biology Centre (QBiC) at the University of Tuebingen for whole genome sequencing.

Sequence data were generated on an Illumina HiSeq 2500 instrument (chemistry SBS v3 plus ClusterKit cBot HS) and processed as described previously [49] but with minor modifications that follow. Supplied sequence data were checked using fastQC v0.11.5 [50]. Data were trimmed with Trim Galore! (--clip_R1 10 --clip_R2 10 --three_prime_clip_R1 10 --three_prime_clip_R2 10 --length 50; Babraham Bioinformatics). Mouse DNA within samples was detected by mapping reads against the mouse genome (GRCm38). Mouse-filtered read files (with an average of 3.58 ± 0.08 Gb sequence data per sample) were used for all subsequent analyses. Kraken2 2.0.8-beta [51] with the pre-compiled Genome Taxonomy Database [52] Kraken2 GTDB_r89_54k index (downloaded on 3 May 2020) available from https://bridges.monash.edu/ndownloader/files/16378439 [53] was used to determine the bacterial and archaeal taxonomic composition/abundance for each sample. Functional annotation was achieved by mapping centroid protein sequences generated as described before [49,51] using the eggNOG-mapper software (v.1.0.3) [54] and associated database (v.4.5). Microbial gene richness was determined as previously described [49]. Data were downsized to adjust for sequencing depth and technical variability by randomly selecting 20 million reads mapped to the merged gene catalogue (of 1,540,712 genes) for each sample and then computing the mean number of genes over 30 random drawings.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Sample treatment before protein extraction

Mouse faecal pellets obtained from wild-type B6EiC3SnF1/J mice were used to compare sample initial preparation methodologies. For the LSC procedure, faeces (~50 mg) were resuspended in phosphate buffer (50 mM Na_2HPO_4/NaH_2PO_4, pH 8.0, 0.1 % Tween 20, 35x volume per mg) by vortexing vigorously for 5 min using 4 mm glass beads (ColiRollers™ Plating beads, Novagen), followed by incubation in a sonication bath for 10 min and shaking at 1,200 rpm for 10 min in a Thermomixer with a thermo block for reaction tubes. Insoluble material was removed by centrifugation at 200 × g at 4 °C for 15 min. The supernatant was removed and the remaining pellet was subjected to two additional rounds of microbial cell extraction. After merging supernatants, microbial cells were collected by centrifugation at 13,000 × g at 4 °C for 30 min. The pellet was resuspended in 80 µL sodium dodecyl sulfate (SDS) buffer (2 % SDS, 20 mM Tris, pH 7.5; namely pellet extraction buffer) and heated at 95 °C for 30 min in a Thermomixer.

For the nLSC procedure, mouse faeces (~25 mg) were homogenised in 150 µL pellet extraction buffer as described above with the following changes. A bead mixture of 0.1 mm glass beads (100 mg), 5 × 1.4 mm ceramic beads (Biolab products), and 1 × 4 mm glass bead was used for five cycles of homogenisation.

Cell lysis and protein extraction

To compare protein extraction by bead beating, two bead beaters were used to disrupt bacterial cell pellets derived from LSC or nLSC procedures. Samples were split in two and were homogenised using 0.1 mm glass beads (100 mg, Sartorius™ Glass Beads) together with the FastPrep-24 5G instrument (MP) at 4 m/s or the BeadBug microtube homogeniser (BeadBug) at 4,000 rpm. Three homogenisation cycles were performed and consisted of 1 min bead beating, 30 sec incubation at 95 °C and 30 sec centrifugation at 13,000 × g. The homogenate was diluted with 800 µL $MgCl_2$ buffer (0.1 mg/mL $MgCl_2$, 50 mM Tris, pH 7.5) and

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

522 centrifuged at 13,000 rpm for 15 min. Proteins from the supernatant were precipitated
523 overnight in acetone/methanol at -20 °C (acetone:methanol:sample with 8:1:1 ratio). Protein
524 pellets were resuspended in 120 µL denaturation buffer (6 M urea, 2 M thiourea, 10 mM Tris,
525 pH 8.0) for downstream use.
526 Additional protein extraction methods were compared only using three biological samples that
527 were subjected to the LSC procedure (as described above). Microbial pellets were prepared in
528 order to compare (1) bead beating, (2) ultrasonication and (3) grinding on liquid
529 nitrogen/ultrasonication. Bead beating was performed as described above by resuspension in
530 100 µL pellet extraction buffer and by using the MP bead beater. The microbial pellets for
531 ultrasonication were resuspended in 120 µL pellet extraction buffer and incubated at 60 °C and
532 1,400 rpm for 10 min. After addition of 1 mL pellet extraction buffer, benzonase was added
533 for DNA removal (final concentration of 1 µL/mL). Ultrasonication was performed using an
534 ultrasonicator with an amplitude of 50 % and a cycle time of 0.5 for 2 min on ice. The samples
535 were incubated at 37 °C and 1,400 rpm for 10 min, followed by centrifugation at 4 °C and
536 10,000 × g for 15 min. The third procedure included grinding of the microbial pellet in liquid
537 nitrogen for 1-2 min with a pestle in reaction tube, followed by ultrasonication as described
538 above.

539 Protein digestion

540 Following extraction, protein amount was quantified using Bradford assay (Bio-Rad, Munich,
541 Germany) [55] and two methods were compared to digest proteins extracted from LSC or nLSC
542 procedures.
543 The in-solution digestion method was performed as follows. Proteins (20 µg starting material)
544 were reduced in 1 mM dithiothreitol (DTT) and alkylated in 5.5 mM iodoacetamide at room
545 temperature (RT) for 1 h each. Proteins were pre-digested with LysC at RT for 3 h using a
546 protein to protease ratio of 75:1. Samples were diluted nine-fold with 50 mM ammonium

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

bicarbonate and digested overnight with trypsin (Sequencing Grade Modified Trypsin, Promega) at pH 8.0 using a protein to protease ratio of 75:1.

Filter-aided sample preparation (FASP) was performed as previously published [56]. Briefly, proteins (10 µg starting material) were reduced in 0.1 M DTT for 40 min at RT. The reduced samples were added to the filter units (30 kDa membrane cut off) and centrifuged at 14,000 × g for 15 min. All further centrifugation steps were performed similarly unless otherwise noted. Samples were then washed with 2X 200 µL urea buffer (100mM Tris/HCl, pH 8.5, 8M urea) and centrifuged. Proteins were incubated in 50 mM IAA for 20 min at RT in the dark. After alkylation, samples were centrifuged and washed three times with 100 µL urea buffer. This was followed by three wash steps with 50 mM ammonium bicarbonate (ABC) for 10 min. Proteins were digested overnight at 37 °C using trypsin digestion (Sequencing Grade Modified Trypsin, Promega) at pH 8.0 using a protein to protease ratio of 100:1. On the following day, the peptides were centrifuged into fresh tubes at 14,000 × g for 10 min. An additional 40 µL ABC buffer was added to the filter units and this solution was also centrifuged to increase the peptide yield. Following digestion either in-solution or FASP, samples were acidified to pH 2.5 with formic acid and cleaned for LC-MS/MS measurement using Empore C18 disks in StageTips [57].

LC-MS/MS measurements

Samples were measured on an EASY-nLC 1200 (Thermo Fisher Scientific) coupled to a Q Exactive HF mass spectrometer (Thermo Fisher Scientific). Peptides were chromatographically separated using 75 µm (ID), 20 cm packed in-house with reversed-phase ReproSil-Pur 120 C18-AQ 1.9 µm resin (Dr. Maisch GmbH).

Peptide samples generated as part of the laboratory method optimisation were eluted over 43 min using a 10 to 33 % gradient of solvent B (80 % ACN in 0.1 % formic acid) followed by a washout procedure. Peptide samples generated as part of the data analysis assessment were

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

eluted over 113 min using a 10 to 33 % gradient of solvent B (80 % ACN in 0.1 % formic acid) followed by a washout procedure.

MS1 spectra were acquired between 300-1,650 Thompson at a resolution of 60,000 with an AGC target of 3×10^6 within 25 ms. Using a dynamic exclusion window of 30 sec, the top 12 most intense ions were selected for HCD fragmentation with an NCE of 27. MS2 spectra were acquired at a resolution of 30,000 and a minimum AGC of 4.5×10^3 within 45 ms.

LC-MS/MS data processing

Raw data obtained from the instrument were processed using MaxQuant (version 1.5.2.8) [58]. The protein sequence databases used for database search consisted of the complete *Mus musculus* Uniprot database (54,506 sequences) and frequently observed contaminants (248 entries), as well as the mouse microbiome catalogue (~2.6 million proteins) [26] for the raw data from laboratory method optimisation samples or the matching metagenome gene translation (~1.5 million proteins) for the raw data from data analysis assessment samples. A FDR of 1 % was required at the peptide and protein levels. A maximum of two missed cleavages was allowed and full tryptic enzyme specificity was required. Carbamidomethylation of cysteines was defined as fixed modification, while methionine oxidation and N-terminal acetylation were set as variable modifications. Match between runs was enabled where applicable. Quantification was performed using label-free quantification (LFQ) [59] and a minimum ratio of 1. All other parameters were left to MaxQuant default settings.

Comparison of sample preparation methods

Unless stated otherwise, the analyses described below were performed in the R environment [60]. To compare the different centrifugation, digestion and lysis methods, we counted for each sample the number of peptide and protein groups with intensities and LFQ intensities superior to zero, respectively. We tested for significant differences between methods using paired t-tests

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

via the ggplot2 package [61]. Quantified peptides and protein groups were checked for overlap between the centrifugation methods using the VennDiagram package. The proportion of host (*Mus musculus*) proteins was computed by summing up all host proteins iBAQ values and then dividing by the total iBAQ per sample. The centrifugation methods were evaluated using a paired t-test.

The taxonomy representation for the centrifugation methods was done via the Unipept software [24]. The quantified peptides (intensity superior to zero) were imported into Unipept with I-L not equal and advanced missed cleavages handling selected only for peptides with 1 or 2 miss-cleavages. The Unipept result were used to count the number of non-redundant peptides assigned to each taxonomic node. The Firmicutes to Bacteroidetes ratio was calculated by summing the spectral count for each phylum and sample. The centrifugation methods were compared on this basis using a paired t-test.

For the differential protein abundance analysis (between LSC and nLSC), the MSnBase package was used as organisational framework for the protein groups LFQ data [62]. Host proteins, reverse hit and potential contaminant proteins were filtered out. Protein groups were retained for further analysis only if more than 90 % of samples within either LSC or nLSC group had an LFQ superior to the first quartile overall LFQ. Significantly changing proteins were identified using paired t-test. Significance was set at an adjusted p-value of 0.01 following Benjamini-Hochberg multiple correction testing, as well as a minimum LSC/nLSC fold-change of ±1.5. The over-representation and GSEA testing of KEGG pathways were done for the significantly up- and down-regulated proteins as well as for the proteins uniquely identified per group via the clusterProfiler package based on hypergeometric distribution (p-adj. ≤ 0.05) [63]. Selected over-represented KEGG pathways were displayed in context of protein quantification using the pathview package.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

MS instruments comparison

The samples generated as part of the data analysis assessment were measured on a Q Exactive HF mass spectrometer as described above, as well as on an Orbitrap Elite mass spectrometer, as described below.

Samples were measured on an EASY-nLC 1200 (Thermo Fisher Scientific) coupled to an Orbitrap Elite mass spectrometer (Thermo Fisher Scientific). Peptides were chromatographically separated using 75 µm (ID), 20 cm packed in-house with reversed-phase ReproSil-Pur 120 C18-AQ 1.9 µm resin (Dr. Maisch GmbH).

Peptide samples were eluted over 213 min using a 10 to 33 % gradient of solvent B (80 % ACN in 0.1 % formic acid) followed by a washout procedure. An optimised gradient was also tested whereby peptide samples were eluted over 225 min using a 5 % (0 min), 10 % (5 min), 13 % (80 min), 15 % (170 min) and 30 % (225 min) gradient of solvent B (80 % ACN in 0.1 % formic acid) followed by a washout procedure.

MS1 spectra were acquired between 300-2,000 Thompson at a resolution of 120,000 with an AGC target of 1×10^6 within 100 ms. Using a dynamic exclusion window of 30 sec, the top 15 most intense ions were selected for HCD fragmentation with an NCE of 35. MS2 spectra were acquired at a resolution of 120,000 and a minimum AGC of 5×10^3 within 150 ms.

Raw data were processed, together with raw data measured on Q Exactive mass spectrometer, as described in the LC-MS/MS data processing section (the matching metagenome gene translation (1,595,268 entries) was used as microbial database). To compare the different measurement methods, we counted for each sample the number of identified MS/MS, non-redundant peptides and protein groups. These were then tested for significant differences between measurement methods using paired t-tests via the ggplot2 package [61].

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Single- versus two-step assessment

HeLa cells were prepared for LC-MS/MS measurements using published method [64]. Briefly, cells were grown in DMEM medium and harvested at 80 % confluence. Proteins were precipitated using acetone and methanol. Proteins were reduced with DTT and digested with Lys-C and trypsin. Peptides were purified on Sep-Pak C18 Cartridge.

Sample was measured as described in the LC-MS/MS measurements section but for a few changes. Peptide sample was eluted over 213 min using a 7 % (0 min), 15 % (140 min) and 33 % (213 min) gradient of solvent B (80 % ACN in 0.1 % formic acid) followed by a washout procedure. The top 10 most intense ions were selected for HCD fragmentation.

Raw data were processed using MaxQuant (version 1.5.2.8) [58] as described in the LC-MS/MS data processing section with a few alterations. The protein sequence databases used for database search consisted of the complete *Homo sapiens* Uniprot database (93,799 sequences) and frequently observed contaminants (248 entries), as well as the mouse microbiome catalogue (~2.6 million proteins) [26]. Several processings were performed differing in the number of microbiome catalogue entries included, such as 1:0, 1:1, 1:2, 1:5, 1:10 and 1:20 *H. sapiens*:bacteria protein sequences and using a single- or two-step database search [25].

Identified MS/MS, peptides and protein groups were assigned to kingdom of origin (conflicts were resolved to Eukaryota by default). To compare the different database search strategies, we counted the number of identified MS/MS, non-redundant peptides and protein groups associated to each kingdom (as well as reverse hits and potential contaminants). We also calculated the FDR based solely on reverse hits or together with bacterial hits in order to investigate the true number of false positives.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Microbiome protein sequences databases comparison

The samples generated as part of the data analysis assessment were measured on a Q Exactive HF mass spectrometer as described above, but were processed using additional databases and search strategies in MaxQuant. Only the microbiome sequence databases differed and consisted of one of (1) UniProt bacterial reference proteomes (5,408,622 protein entries), (2) UniProt bacterial pan proteomes (18,541,701 entries), (3) protein sequences from the Mouse Gut Metagenome catalogue (2,626,630 entries) [26], or (4) the matching metagenome gene translation (1,595,268 entries). Processing also involved comparison between single- and two-step database search [25]. To compare the different databases and search strategies, we counted the number of sequences and OTUs per database, as well as reported the identification rates, non-redundant peptides count and protein groups count.

Metagenome to metaproteome correlation

All subsequent sections use the samples generated as part of the data analysis assessment that were measured on a Q Exactive HF mass spectrometer and processed against the matching metagenome gene translation as described above. For direct comparison between metagenome and metaproteome, the identified genes were collapsed into groups identical to protein groups composition from mass spectrometry. Each gene groups abundance was calculated as the highest gene abundance within that group. Each gene groups and corresponding protein groups abundances were correlated across samples using Spearman's rank correlation from the stats package. Significance was set at an adjusted p-value of 0.05 following Benjamini-Hochberg multiple correction testing. The GSEA testing of KEGG pathways and Gene ontologies were performed via the clusterProfiler package based on hypergeometric distribution (p-adj. ≤ 0.05) [63] following z-scoring of Spearman rho estimate per KEGG orthologies.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Metaproteome-based taxonomic representation

Metaproteome-derived taxonomic assignments were obtained using (1) Diamond (v. 0.9.23) [28], (2) Unipept online (v. Dec. 2018) [23], or (3) Kraken2 (2.0.8-beta) [51] softwares for either the identified peptides or all matched-metagenome protein sequences. The Diamond alignment was performed against NCBI non-redundant protein sequences database using sensitive and taxonomic classification mode. The Unipept online analysis was done via the metaproteome analysis function with I-L not equal and advanced missed cleavages handling selected against all UniProt entries. The Kraken2 k-mer analysis was carried out at the nucleotide level (corresponding to identified peptides or proteins) against the Genome Taxonomy Database (v. 89) [65] obtained from Struo software [66]. For each software approach, the complete taxonomic lineage (NCBI or GTDB) was retrieved per peptide or protein groups and the lowest common ancestor was determined. OTUs were quantified per sample based on the different software approaches by summing either (1) PSM count, (2) peptide intensity, (3) protein groups iBAQ, (4) protein groups intensity, (5) protein groups LFQ, (6) protein groups MS/MS count. OTUs quantification were normalised on a per sample basis as percentage of total to get OTUs representation. OTUs were filtered on a per sample basis based on minimum representation threshold (representation $\geq$ X, with X equal from 0 % to 10 %), while the correlation between metagenome and metaproteome was calculated using Spearman's rank correlation. The optimal representation filtering threshold was identified as the threshold that maximises number of identified OTUs and correlation between omics (i.e. count × spearman rho). Using the optimal representation threshold and quantification approach, the overlap in identified OTUs between omics was calculated, as well as the Spearman's rank correlation at each taxonomic level.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Functional KEGG categories representation

For each sample, the protein groups iBAQ values were summed per KEGG category (level 2) on the basis of KEGG orthology annotation. The same approach was also undertaken for gene count. The KEGG category abundance were normalised for differing number of KO entries per category and for variation between samples; this was done separately for metagenome and metaproteome. Differences in KEGG category abundance between metagenome and metaproteome were tested using paired t-tests from the stats package. Significance was set at an adjusted p-value of 0.01 following Benjamini-Hochberg multiple correction testing. Significantly changing KEGG categories were prioritised based on gene groups to protein groups correlation (see section Metagenome to metaproteome correlation), whereby the Wilcoxon rank-sum test was used to identify KEGG category containing KO entries whose correlation differ from overall distribution (adjusted p-value ≤ 0.05).

To investigate further these selected KEGG categories, the protein groups iBAQ and gene count were used as described in the previous paragraph to derive KO normalised abundance and t-test results. Using the KO entries from each selected KEGG categories, separate GSEA testing of KEGG pathways were performed via the clusterProfiler package based on hypergeometric distribution (p-adj. ≤ 0.05).

Acknowledgments

MED, BM, XA and LD are grateful to the European Community 7th Framework Program under Coordinated Action NEURON-ERANET (grant agreement 291840). BM was supported by grants from the Deutsche Forschungsgemeinschaft (German Research Foundation Cluster of Excellence EXC 2124). BM and NN acknowledge support by the High Performance and Cloud Computing Group at the Center for Data Processing of the University of Tübingen, the state of Baden-Wuerttemberg through bwHPC. The metagenomic work detailed herein used the

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

computing resources of the UK MEDical BIOinformatics partnership – aggregation, integration, visualization and analysis of large, complex data (UK Med-Bio) – which was supported by the Medical Research Council (grant number MR/L01632X/1). XA acknowledge support from the MINECO, Spain (grant number PCIN-2014-105).

Author contributions

LD, XA, MD and BM designed the study. XA and CG generated the mouse cohorts and collected the murine faecal material. VA, TG and ID prepared the murine faecal samples for proteomic measurement by mass spectrometry. LH processed the metagenomic data, generating the taxonomic and gene abundance outputs. NN processed the metaproteomic datasets and performed the proteogenomic integration. NN wrote the manuscript with the input from all authors.

Data Access

The complete metaproteomic bioinformatic workflow is available online [67]. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [68] partner repository with the dataset identifiers PXD020695, PXD020738, PXD021928 and PXD021932. Trimmed whole genome sequence data with mouse reads removed have been deposited with GenBank, EMBL and DDBJ databases under the BioProject accession PRJNA473429.

References

1. Jandhyala SM, Talukdar R, Subramanyam C, et al. Role of the normal gut microbiota. World J Gastroenterol. 2015;21(29):8787-8803.
2. Valdes AM, Walter J, Segal E, et al. Role of the gut microbiota in nutrition and health. BMJ. 2018;361:k2179.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

3. Arumugam M, Raes J, Pelletier E, et al. Enterotypes of the human gut microbiome. Nature. 2011 2011/05//;473(7346):174-180.
4. Vieira-Silva S, Falony G, Belda E, et al. Statin therapy is associated with lower prevalence of gut microbiota dysbiosis. Nature. 2020 May;581(7808):310-315.
5. Wang J, Linnenbrink M, Künzel S, et al. Dietary history contributes to enterotype-like clustering and functional metagenomic content in the intestinal microbiome of wild mice. Proceedings of the National Academy of Sciences. 2014;111(26):E2703.
6. Ladau J, Eloe-Fadrosh EA. Spatial, Temporal, and Phylogenetic Scales of Microbial Ecology. Trends in Microbiology. 2019 2019/08/01/;27(8):662-669.
7. Parfrey LW, Knight R. Spatial and temporal variability of the human microbiota. Clinical Microbiology and Infection. 2012 2012/07/01/;18:5-7.
8. Sampson TR, Debelius JW, Thron T, et al. Gut Microbiota Regulate Motor Deficits and Neuroinflammation in a Model of Parkinson's Disease. Cell. 2016 Dec 1;167(6):1469-1480 e12.
9. Ley RE, Turnbaugh PJ, Klein S, et al. Microbial ecology: human gut microbes associated with obesity. Nature. 2006 Dec 21;444(7122):1022-3.
10. Xiao L, Sonne SB, Feng Q, et al. High-fat feeding rather than obesity drives taxonomical and functional changes in the gut microbiota in mice. Microbiome. 2017 Apr 8;5(1):43.
11. Zhang X, Deeke SA, Ning Z, et al. Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. Nat Commun. 2018 Jul 20;9(1):2873.
12. Hsiao EY, McBride SW, Hsien S, et al. Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. Cell. 2013 Dec 19;155(7):1451-63.
13. Tengeler AC, Dam SA, Wiesmann M, et al. Gut microbiota from persons with attention-deficit/hyperactivity disorder affects the brain in mice. Microbiome. 2020 2020/04/01;8(1):44.
14. Wang L, Christophersen CT, Sorich MJ, et al. Increased abundance of Sutterella spp. and Ruminococcus torques in feces of children with autism spectrum disorder. Mol Autism. 2013 Nov 4;4(1):42.
15. Biagi E, Candela M, Centanni M, et al. Gut Microbiome in Down Syndrome. PLOS ONE. 2014;9(11):e112023.
16. Hettich RL, Pan C, Chourey K, et al. Metaproteomics: harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communities. Anal Chem. 2013 May 7;85(9):4203-14.
17. Li X, LeBlanc J, Truong A, et al. A metaproteomic approach to study human-microbial ecosystems at the mucosal luminal interface. PloS one. 2011;6(11):e26542-e26542.
18. Ram RJ, Verberkmoes NC, Thelen MP, et al. Community proteomics of a natural microbial biofilm. Science. 2005 Jun 24;308(5730):1915-20.
19. Tanca A, Palomba A, Pisanu S, et al. Enrichment or depletion? The impact of stool pretreatment on metaproteomic characterization of the human gut microbiota. Proteomics. 2015 Oct;15(20):3474-85.
20. Zhang X, Li L, Mayne J, et al. Assessing the impact of protein extraction methods for human gut metaproteomics. J Proteomics. 2018 May 30;180:120-127.
21. Muth T, Behne A, Heyer R, et al. The MetaProteomeAnalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation. J Proteome Res. 2015 Mar 6;14(3):1557-65.
22. Cheng K, Ning Z, Zhang X, et al. MetaLab: an automated pipeline for metaproteomic data analysis. Microbiome. 2017 Dec 2;5(1):157.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

23. Mesuere B, Devreese B, Debyser G, et al. Unipept: tryptic peptide-based biodiversity analysis of metaproteome samples. J Proteome Res. 2012 Dec 7;11(12):5773-80.
24. Mesuere B, Van der Jeugt F, Willems T, et al. High-throughput metaproteomics data analysis with Unipept: A tutorial. J Proteomics. 2018 Jan 16;171:11-22.
25. Jagtap P, Goslinga J, Kooren JA, et al. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. Proteomics. 2013 Apr;13(8):1352-7.
26. Xiao L, Feng Q, Liang S, et al. A catalog of the mouse gut metagenome. Nature biotechnology. 2015 Oct;33(10):1103-8.
27. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014 Mar 3;15(3):R46.
28. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015 Jan;12(1):59-60.
29. Kleiner M, Thorson E, Sharp CE, et al. Assessing species biomass contributions in microbial communities via metaproteomics. Nat Commun. 2017 Nov 16;8(1):1558.
30. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000 Jan 1;28(1):27-30.
31. Zhang X, Ning Z, Mayne J, et al. MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. Microbiome. 2016 Jun 24;4(1):31.
32. Tanca A, Manghina V, Fraumene C, et al. Metaproteogenomics Reveals Taxonomic and Functional Changes between Cecal and Fecal Microbiota in Mouse. Front Microbiol. 2017;8:391.
33. Michalski A, Damoc E, Hauschild JP, et al. Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. Mol Cell Proteomics. 2011 Sep;10(9):M111 011015.
34. Williamson JC, Edwards AV, Verano-Braga T, et al. High-performance hybrid Orbitrap mass spectrometers for quantitative proteome analysis: Observations and implications. Proteomics. 2016 Mar;16(6):907-14.
35. Shishkova E, Hebert Alexander S, Coon Joshua J. Now, More Than Ever, Proteomics Needs Better Chromatography. Cell Systems. 2016 2016/10/26/;3(4):321-324.
36. Nguyen TLA, Vieira-Silva S, Liston A, et al. How informative is the mouse for human gut microbiota research? Disease Models & Mechanisms. 2015;8(1):1.
37. Hart ML, Meyer A, Johnson PJ, et al. Comparative Evaluation of DNA Extraction Methods from Feces of Multiple Host Species for Downstream Next-Generation Sequencing. PLOS ONE. 2015;10(11):e0143334.
38. Nagpal R, Wang S, Solberg Woods LC, et al. Comparative Microbiome Signatures and Short-Chain Fatty Acids in Mouse, Rat, Non-human Primate, and Human Feces. Front Microbiol. 2018;9:2897-2897.
39. Speicher KD, Kolbas O, Harper S, et al. Systematic analysis of peptide recoveries from in-gel digestions for protein identifications in proteome studies. J Biomol Tech. 2000;11(2):74-86.
40. Muth T, Kolmeder CA, Salojarvi J, et al. Navigating through metaproteomics data: A logbook of database searching. Proteomics. 2015 Oct;15(20):3439-53.
41. Heyer R, Schallert K, Zoun R, et al. Challenges and perspectives of metaproteomic data analysis. J Biotechnol. 2017 Nov 10;261:24-36.
42. Li J, Jia H, Cai X, et al. An integrated catalog of reference genes in the human gut microbiome. Nature biotechnology. 2014 2014/08/01;32(8):834-841.
43. Chen T, Yu WH, Izard J, et al. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. Database (Oxford). 2010 Jul 6;2010:baq013.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

44. Grassl N, Kulak NA, Pichler G, et al. Ultra-deep and quantitative saliva proteome reveals dynamics of the oral microbiome. Genome medicine. 2016;8(1):44.
45. Erickson AR, Cantarel BL, Lamendella R, et al. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. PLoS One. 2012;7(11):e49138.
46. Mills RH, Vazquez-Baeza Y, Zhu Q, et al. Evaluating Metagenomic Prediction of the Metaproteome in a 4.5-Year Study of a Patient with Crohn's Disease. mSystems. 2019 Jan-Feb;4(1).
47. Edirisinghe JN, Weisenhorn P, Conrad N, et al. Modeling central metabolism and energy biosynthesis across microbial life. BMC Genomics. 2016 Aug 8;17:568.
48. Sidoli S, Kulej K, Garcia BA. Why proteomics is not the new genomics and the future of mass spectrometry in cell biology. J Cell Biol. 2017 Jan 2;216(1):21-24.
49. Hoyles L, Fernandez-Real JM, Federici M, et al. Molecular phenomics and metagenomics of hepatic steatosis in non-diabetic obese women. Nat Med. 2018 Jul;24(7):1070-1080.
50. Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/; 2010.
51. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019 Nov 28;20(1):257.
52. Parks DH, Chuvochina M, Chaumeil PA, et al. A complete domain-to-species taxonomy for Bacteria and Archaea. Nature biotechnology. 2020 Sep;38(9):1079-1086.
53. Méric G, Wick RR, Watts SC, et al. Correcting index databases improves metagenomic studies. bioRxiv. 2019:712166.
54. Huerta-Cepas J, Forslund K, Coelho LP, et al. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Molecular Biology and Evolution. 2017;34(8):2115-2122.
55. Bradford MM. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem. 1976 May 7;72:248-54.
56. Wisniewski JR, Zougman A, Nagaraj N, et al. Universal sample preparation method for proteome analysis. Nat Methods. 2009 May;6(5):359-62.
57. Rappsilber J, Mann M, Ishihama Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat Protoc. 2007;2(8):1896-906.
58. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nature biotechnology. 2008 Dec;26(12):1367-72.
59. Cox J, Hein MY, Luber CA, et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol Cell Proteomics. 2014 Sep;13(9):2513-26.
60. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2018.
61. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2016.
62. Gatto L, Lilley KS. MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. Bioinformatics. 2012 Jan 15;28(2):288-9.
63. Yu G, Wang LG, Han Y, et al. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012 May;16(5):284-7.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

64. Schmitt M, Sinnberg T, Nalpas NC, et al. Quantitative Proteomics Links the Intermediate Filament Nestin to Resistance to Targeted BRAF Inhibition in Melanoma Cells. Mol Cell Proteomics. 2019 Jun;18(6):1096-1109.
65. Parks DH, Chuvochina M, Waite DW, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature biotechnology. 2018 Nov;36(10):996-1004.
66. de la Cuesta-Zuluaga J, Ley RE, Youngblut ND. Struo: a pipeline for building custom databases for common metagenome profilers. Bioinformatics. 2020 Apr 1;36(7):2314-2315.
67. Nalpas N, Macek B. Integrated metaproteomics workflow. 1.0. Zenodo; 2020.
68. Perez-Riverol Y, Csordas A, Bai J, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res. 2019 Jan 8;47(D1):D442-D450.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Tables

Table 1: Performance comparison of different sample preparation and data analysis steps. In bold are the best methods according to assessed criteria: peptide/protein count, host/dietary contamination, *Firmicutes* to *Bacteroidetes* ratio, time efficiency, sample amount, FDR, identification rate, OTU assigned peptides and number of OTUs. The performance status is displayed using minus sign for poor, equal sign for similar/no difference or plus sign for good performance.

		Peptide/protein count	Host/dietary contamination	Firmicutes:Bacteroidetes	Time efficiency	Sample amount	FDR	Identification rate	OTU assigned peptides	Number of OTUs
Centrifugation	**LSC**	+	-	+	-	=				
	nLSC	-	+	-	+	=				
Digestion	**In-solution**	+			-	=				
	FASP	-			+	=				
Lysis	**Bead-beating**	=			+	=				
	Nitrogen	=			-	=				
	Sonication	=			-	=				
MS instrument	Elite	-			-	-				
	Elite long LC	-			-	-				
	Q Exactive	++			+	+				
Search strategy	**Single-step**	-			+		+	-		
	Two-step	+			-		--	+		
Protein sequence database	UP Ref proteomes	--			-		-	--		
	UP Pan proteomes	--			-		-	--		
	Catalogue	+			+		+	+		
	Metagenome	++			+		+	++		
OTUs quantification	Diamond				--				--	-
	Unipept				+				-	-
	Kraken2				-				+	++

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figures

Figure 1: Low speed centrifugation impacts protein identification and taxonomic representation. A) Number of identified peptides and protein groups per samples for the comparison between LSC (red) and nLSC (blue) methods. B) Number of identified peptides and protein groups per samples for the comparison between LSC-in solution digestion (red), LSC-FASP (grey), nLSC-in solution digestion (blue) and nLSC-FASP (orange) methods. C) Number of identified peptides and protein groups per samples for the comparison between bead-beating (red), nitrogen (grey) and sonication (blue) lysis methods. A-C) Represented significance results correspond to paired t-test on N = 12: * *p*- value ≤ 0.05, ** ≤ 0.01, *** ≤ 0.01. D-E) Unipept-derived taxonomic representation (down to phylum level) for the peptide identified in the LSC (D) and nLSC (E) samples. Number represent the Peptide Spectrum Match count for each taxon.

Figure 2: Functional representation of proteins detected in LSC- and nLSC-samples is similar despite different peptide and protein abundances. A) Overlap in the overall identified peptides or protein groups between the LSC and nLSC methods. B) Principal component analysis showing separation of the samples based on biological replicates and sample preparation (LSC or nLSC) and the clustering of the technical replicates. C) Volcano plot of the protein abundance comparison between LSC and nLSC approaches. Significant protein groups based on paired t-test from N = 12 with FDR ≤ 0.01 and absolute fold-change ≥ 2.5. D) KEGG pathways over-representation testing for the significantly up-regulated (red) and down-regulated (blue) protein groups between LSC and nLSC sample preparation approaches. Fisher exact-test threshold set to adjusted *p*-value ≤ 0.1.

Figure 3: Two-step database search in combination with target-decoy strategy leads to a dramatic increase in false positive rate. A) The protein groups count is shown for single- or two-step search strategies across increasingly large protein sequence databases. Counts are

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

colour-coded per category, with eukaryote (grey), bacteria (red), contaminant (blue) and reverse (orange) hits. B) The FDR is calculated for single- or two-step search strategies across increasingly large protein sequence databases. The FDR is calculated based on reverse hits only (circle shape) or reverse plus bacterial hits (triangle shape). C) The faeces metaproteome data from a cohort of 38 mice were searched against different protein sequence databases and either via single- or two-step search strategy. The comparison between databases focuses on the number of sequences, number of OTUs, identification rate, number of peptides identified and number of protein groups identified. The following databases were compared: UniProt bacterial reference proteome (blue colours), UniProt bacterial pan-proteome (black colours), mouse microbiome catalog (orange) and matching metagenome sequences (red).

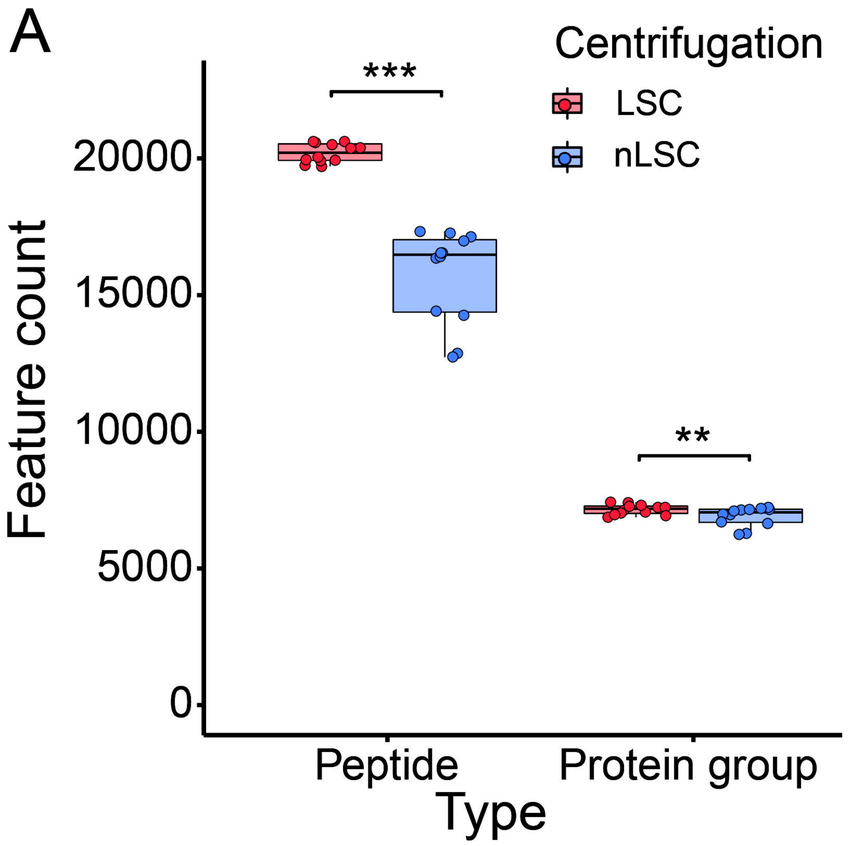
Figure 4: Metaproteome and metagenome show positive correlation at the level of protein, gene and taxonomic representation. A) Assessment of the optimal filtering threshold to maximise the total number of identified OTUs and the OTUs abundance correlation between metaproteome- and metagenome-derived taxonomy representation. For the filtering, OTUs are retained on a per sample basis when having a taxonomic representation superior or equal to defined filtering value. Metaproteome-derived taxonomic representation was calculated from Kraken2 software results using different calculation methods (i.e., peptide and protein quantifications). B) The overlap in OTUs between metaproteome and metagenome at different taxonomic levels. Numbers above bars correspond to the spearman rank correlation between metaproteome- and metagenome-derived OTUs abundance for the corresponding taxonomic level with N = 38: * *p*- value $\leq$ 0.05, ** $\leq$ 0.01, *** $\leq$ 0.01. C) Correlation in abundances is shown between each protein groups (metaproteome) and corresponding gene "groups" (metagenome). Correlation was tested using Spearman's rank correlation and *p*-value was adjusted for multiple testing using Benjamini-hochberg correction. Significantly correlating protein/gene groups are in red colours, while significantly anti-correlating

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

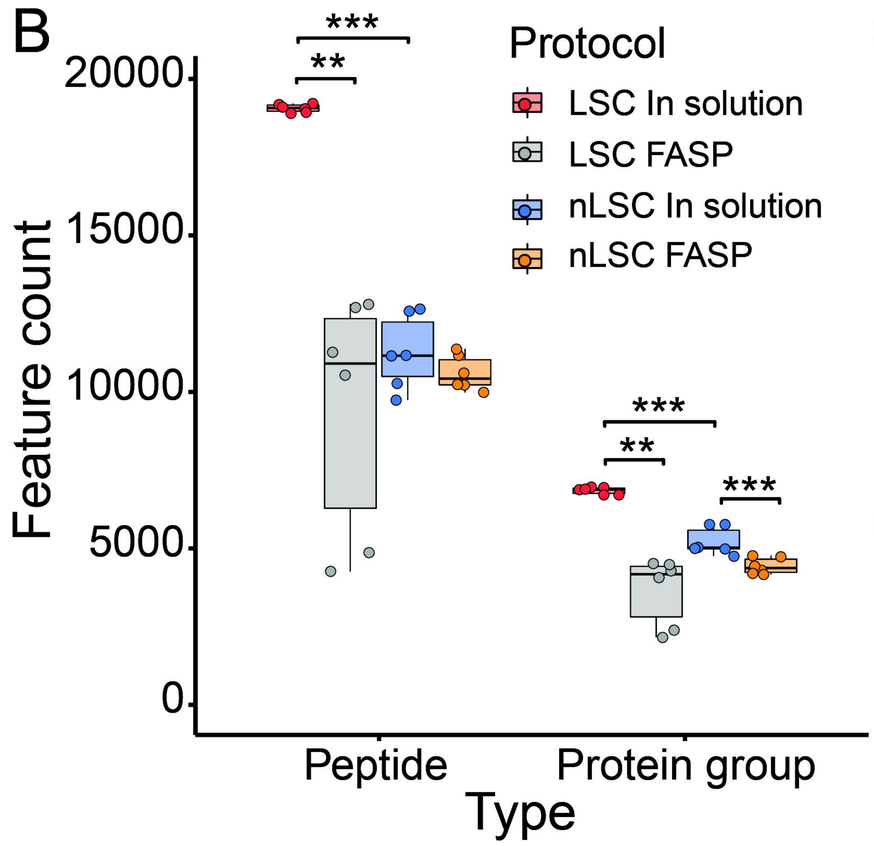
protein/gene groups are in green colours. D) GSEA of KEGG pathways based on ranking of the protein/gene groups correlation. Pathway node colour corresponds to GSEA results adjusted *p*-value and node size matches the number of protein/gene group assigned to the pathway.

Figure 5: Functionally active pathways derived from the metaproteome differs from the metagenome potential. A) Comparison in the proportion of selected KEGG functional categories (level 2) between metaproteome (red) and metagenome (grey). Paired t-test p-values are indicated (N = 38). B) GSEA of KEGG pathways based on ranking of t-test results from KEGG orthology proportion between metaproteome and metagenome. KEGG pathways are colour-coded based on KEGG functional categories (level 2). Only significantly over-represented KEGG pathways are shown with adjusted *p*-value ≤ 0.05. C) Interaction network between KEGG orthologies and KEGG pathways for the KEGG functional category "Protein families: genetic information processing". Pathway node size corresponds to number of KEGG orthologies associated to it. KEGG orthologies are colour-coded based on directional adjusted *p*-value from the t-test comparison between metaproteome and metagenome.

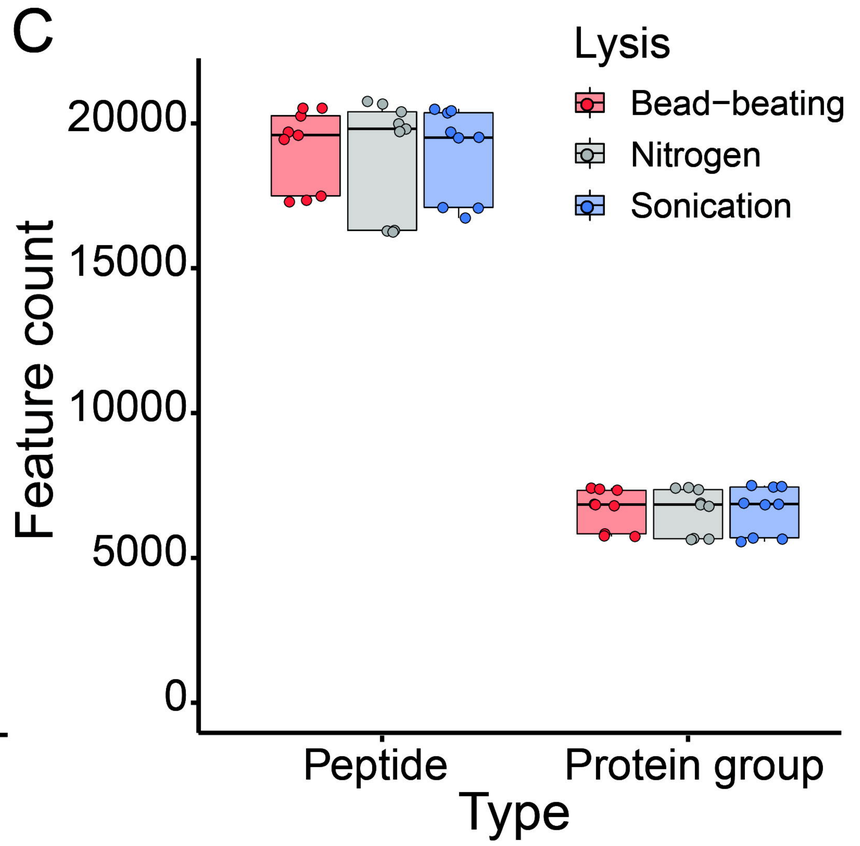
A



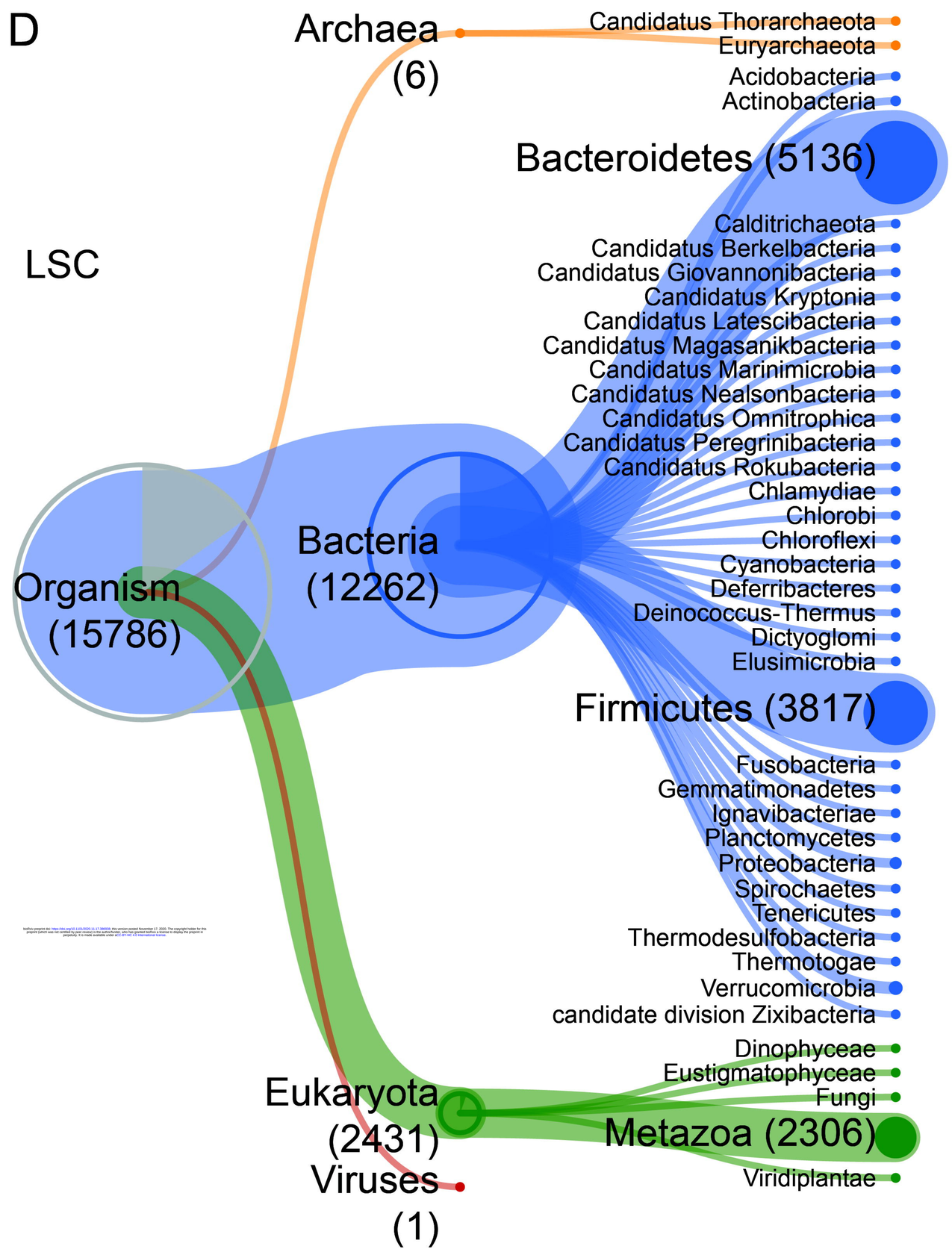
B



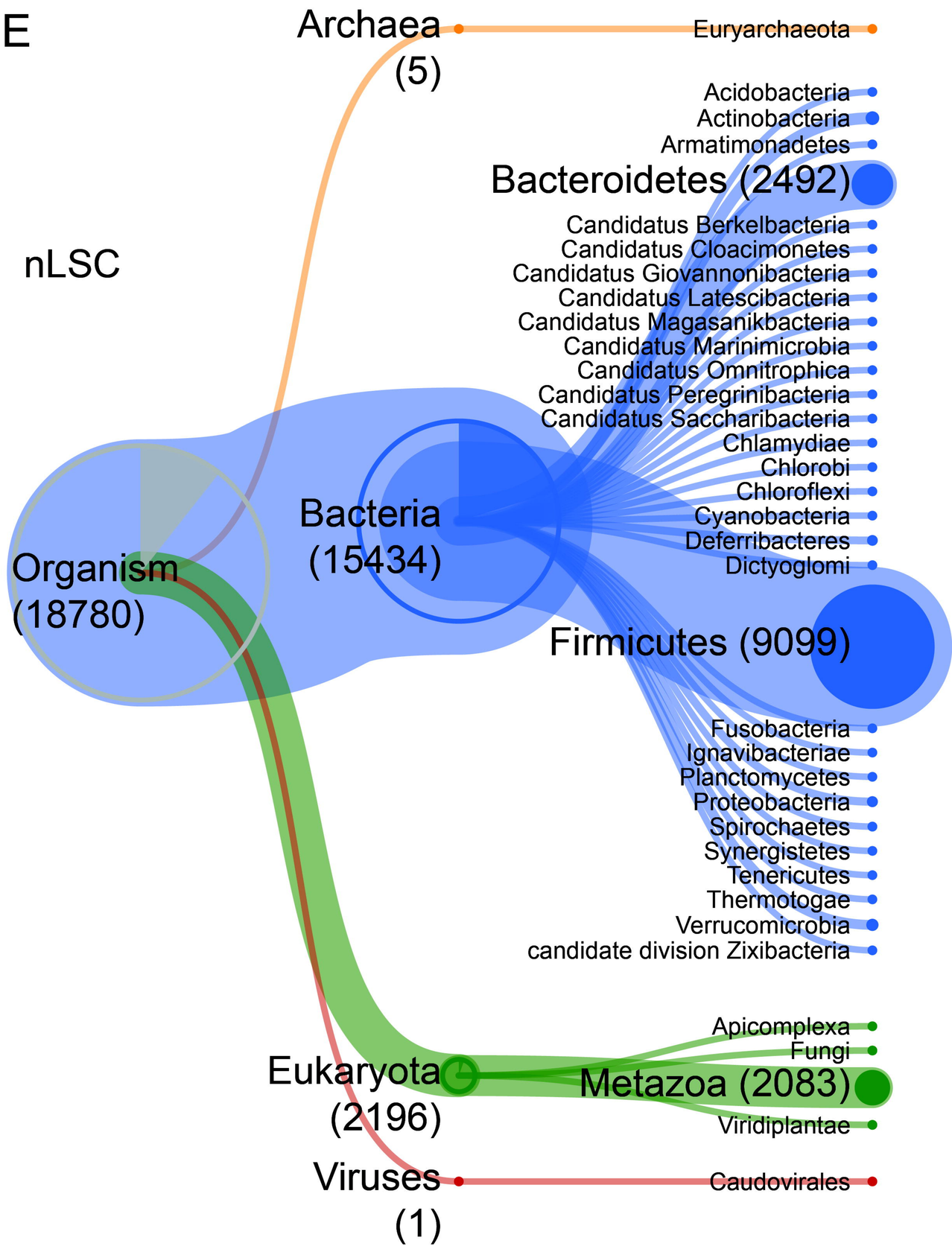
C



D



E



bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A

Peptide

18346 14452 19333

LSC nLSC

Protein group

3778 5956 5665

B

Centrifugation
nLSC LSC
Sample
1 2 3

Dimension 2 (16.4%)

Dimension 1 (23.2%)

C

Up in nLSC

267 365

588 1369

Up in LSC

ASD GROS TPIA GROL GAP PGK SERC PTB RPLB GLGP METK CHEV HBD AMIA RPSA GROS PGK TUF PTB OXC UREC OXC SDHA

$-\text{Log}_{10}$ FDR

L/nL $\log_2$ fold-change

D

Protein used for over-representation
LSC/nLSC down-regulated
LSC/nLSC up-regulated

Ribosome
Glycolysis / Gluconeogenesis
Carbon metabolism
Biosynthesis of secondary metabolites
Biosynthesis of amino acids
Carbon fixation in photosynthetic organisms
Pentose phosphate pathway
RNA degradation
Microbial metabolism in diverse environments
Ribosome
Carbon fixation in photosynthetic organisms
Carbon metabolism
Glycolysis / Gluconeogenesis
Carbon fixation pathways in prokaryotes
Citrate cycle (TCA cycle)
Valine, leucine and isoleucine degradation
Pyruvate metabolism
Microbial metabolism in diverse environments
RNA degradation
Alanine, aspartate and glutamate metabolism
Glyoxylate and dicarboxylate metabolism
Arginine biosynthesis
Propanoate metabolism
Butanoate metabolism
Protein export
Methane metabolism
Biosynthesis of secondary metabolites

$-\text{Log}_{10}$ adj. p–value

Overrepresentation ratio

(which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

A

Category
Contaminant
Bacteria
Eukaryota
Reverse

Protein groups count
6000
4000
2000
0

+ + + + + +
1:0 1:1 1:2 1:5 1:10 1:20

B

Decoy type
Reverse
Reverse + Bacteria

Search
Single−step
Two−step

FDR
0.3
0.2
0.1

Two-step + + + + + +
Ratio 1:0 1:1 1:2 1:5 1:10 1:20

C

UP−Ref. 1st
UP−Ref. 2nd
UP−Pan. 1st
UP−Pan. 2nd
Catalog
Meta-genome

0 0.5 1 1.5
Sequences in DB (x10^7)

0 5000 10000
OTUs in DB

0 10 20
Identification rate (%)

0 50000 100000
Peptides identified

0 10000 20000
Protein groups identified

(which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

A

OTU Spearman ρ

0.75

0.50

0.25

Method

PSM count

Peptide Intensity

Protein iBAQ

Protein Intensity

Protein LFQ intensity

Protein MS/MS Count

Count identified OTU

1000

100

10

0.0

2.5

5.0

7.5

10.0

Minimum OTU representation (%)

B

Count OTU identified

600

400

200

0

Overlap

Both

Metaproteome

Metagenome

0.84***

0.93***

0.81***

0.78***

0.81***

0.74***

phylum

class

order

family

genus

species

C

-log10 adj. p−value

8

6

4

2

0

−0.5

0.0

0.5

1.0

Correlation Spearman ρ

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.386938; this version posted November 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under CC BY-NC-ND 4.0 International license.

D

Metabolic pathways

Biosynthesis of secondary metabolites

Biosynthesis of antibiotics

Microbial metabolism in diverse environments

Propanoate metabolism

Butanoate metabolism

Carbon metabolism

Starch and sucrose metabolism

Citrate cycle (TCA cycle)

Carbon fixation pathways in prokaryotes

Galactose metabolism

Glycolysis / Gluconeogenesis

Pyruvate metabolism

Ribosome

Methane metabolism

Pentose phosphate pathway

RNA degradation

Carbon fixation in photosynthetic organisms

Set size

200

400

Adj. p-value

0.01

0.03

0.4

0.5

0.6

0.7

Enrichment score

A

Membrane transport
1.3e−09
Translation
5.5e−14
Transcription
p < 2.22e−16
Carbohydrate metabolism
1.3e−05
Drug resistance: antimicrobial
p < 2.22e−16
Protein families:
signaling and cellular processes
p < 2.22e−16
Protein families:
genetic information processing
4.5e−08
0.15
0.10
0.05
0.00
Normalised abundance (%)
Metagenome
Metaproteome

B

KEGG level 2
Translation
Protein families: genetic information processing
Membrane transport
Protein families: signaling and cellular processes
Carbohydrate metabolism
Ribosome
Carbon fixation pathways in prokaryotes
Carbon metabolism
Pentose phosphate pathway
Glycolysis / Gluconeogenesis
Biosynthesis of amino acids
Methane metabolism
Microbial metabolism in diverse environments
Biosynthesis of antibiotics
Carbon fixation in photosynthetic organisms
Biosynthesis of secondary metabolites
Mismatch repair
Amino sugar and nucleotide sugar metabolism
Glycolysis / Gluconeogenesis
Aminoacyl−tRNA biosynthesis
Two−component system
Fructose and mannose metabolism
Carbon metabolism
DNA replication
Homologous recombination
Microbial metabolism in diverse environments
Amino sugar and nucleotide sugar metabolism
Ribosome
Quorum sensing
Amino sugar and nucleotide sugar metabolism
2.5
2.0
1.5
−Log10 adj. p−value
−1.0
−0.5
0.0
0.5
1.0
Enrichment score

C

rpsJ
rplA
rpsR
rplW
rpmD
rpsL
rpsF
rplY
rplK
rplB
rpsI
rplO
rplI
rpsP
rpsT
rplN
rpsE
rplL
rplT
rpsO
rplF
rpsH
rpsK
rpmE
rplR
Ribosome
rpmB
rplE
rpsC
rplX
rpsS
rpsD
rpsG
rplS
rplC
rplD
rpsN
rplJ
rplV
rpsM
rpsQ
rpsB
rpsA
rplQ
rplP
rpmC
rplM
rplU
ruvA
recG
Homologous recombination
polA
recJ
dnaB
polC
dnaE
DNA replication
Mismatch repair
ligA, ligB
rnhA
xseA
mutS2
Directional
-log10 adj. p-value
20
0
−20
Size
5
25
45